LIMITACIONES EN EL USO DE CORPUS DIACRÓNICOS DEL ESPAÑOL. NUEVAS APORTACIONES DESDE EL PROYECTO DE INVESTIGACIÓN POST SCRIPTUM

GAEL VAAMONDE
Universidade de Lisboa

RESUMEN

Este trabajo tiene un doble objetivo. Por un lado, se exponen algunas limitaciones de los corpus diacrónicos informatizados que están actualmente disponibles para la investigación del español. Por otro lado, se da a conocer el proyecto de investigación Post Scriptum, que pretende publicar un conjunto de cartas privadas escritas en español y portugués durante la Edad Moderna. Post Scriptum puede aportar soluciones a las carencias de otros corpus similares, convirtiéndose en un recurso adecuado para la investigación en lingüística histórica y en un complemento idóneo de los grandes corpus existentes.

Palabras clave: corpus diacrónico, anotación de corpus, cartas privadas, escritura cotidiana

ABSTRACT

The purpose of this paper is two-fold. On the one hand, some constraints on the computerized diachronic corpora currently available for Spanish research are explained. On the other hand, the Post Scriptum Project is presented, which aims to publish a collection of private letters written in Spanish and Portuguese along the Early Modern period. Post Scriptum can provide solutions to the shortcomings of other similar tools, making it a suitable resource for research in Historical Linguistics and an ideal complement to the existing large corpora.

Keywords: diachronic corpus, corpus annotation, private letters, ordinary writing

1. INTRODUCCIÓN

Como sucede en otras disciplinas científicas, la lingüística histórica se ha beneficiado en los últimos años de las nuevas tecnologías informáticas. La creación de corpus diacrónicos en formato electrónico ha supuesto un avance indiscutible para la investigación del cambio lingüístico, poniendo al servicio del especialista y de cualquier usuario interesado una gran colección de textos históricos con los que poder trabajar sobre una base empírica. En el caso del español, contamos con dos grandes recursos de acceso gratuito en red: El *Corpus Diacrónico del Español* (CORDE), creado por la Real Academia Española, y el *Corpus del Español* (CdE), creado por Mark Davies.

En sus aspectos fundamentales, ambos macrocorpus son bastante parecidos. Los dos cuentan con una gran cantidad de texto recopilado (250 millones y 100 millones de palabras, respectivamente), los dos incluyen una variedad importante de géneros y los dos abarcan un marco cronológico amplio, desde los inicios del idioma hasta finales del siglo XX. Cabe resaltar que CORDE y CdE proporcionan una fuente primaria de datos todavía inexistente en la actualidad para otras lenguas: estamos hablando de verdaderas herramientas de referencia para la investigación diacrónica del español.

2. LIMITACIONES DE LOS GRANDES CORPUS

Pese al valor indiscutible de estos corpus, es posible detectar en ellos ciertos inconvenientes que pueden dificultar o impedir la extracción de determinados datos. Algunas limitaciones del CORDE y del CdE ya han sido puestas de manifiesto en trabajos anteriores (Nieuwenhuijsen 2009, Enrique-Arias 2012). En este artículo me centraré en cuatro desventajas básicas.

3.1 Texto escrito como fuente de datos

Una limitación forzosa y obvia de todo corpus diacrónico es la carencia de fuentes orales, esto es, de transcripciones directas de la lengua hablada. La investigación histórica se ve forzada a asumir la

palabra escrita como fuente fiable de datos, aun cuando debamos reconocer que todo cambio diacrónico se gesta primariamente en el plano de la oralidad. Esta carencia se puede minimizar dando cabida en el corpus a textos propios de un registro informal, aquellos que contengan expresiones más próximos al uso oral del momento. Sin embargo, no es este el caso de los dos corpus citados en el apartado anterior, en los que se aprecia un predominio de documentos literarios o, en todo caso, de textos que revelan una prosa formal y cuidada.

3.2. Niveles de acceso al documento

Una segunda limitación tiene que ver con los niveles de acceso disponibles para cada documento. Para poder satisfacer las expectativas de todos los que se aproximan al texto, lo ideal es ofrecer siempre varios niveles de acceso: una transcripción paleográfica, pensada para llevar a cabo estudios que operan al nivel gráfico y fonético; una edición normalizada del texto, que facilite la lectura al usuario no iniciado y favorezca la base para acometer estudios de índole gramatical; y una imagen del facsímil, que permita cotejar el original así como corregir lecturas dudosas o erróneas. Ni CORDE ni CdE pueden ser aprovechados en todos estos niveles, puesto que carecen de una transcripción paleográfica uniforme y no ofrecen la consulta del facsímil[1].

3.3. Anotación lingüística detallada

En tercer lugar hay que señalar la falta de anotación lingüística detallada, aunque en este apartado es necesario establecer una clara distinción entre ambos corpus. El CORDE, como proyecto ya cerrado, no está lematizado ni etiquetado, lo que dificulta considerablemente la investigación morfológica y sintáctica. Es cierto que su interfaz permite hacer uso de ciertas expresiones regulares para la búsqueda de subcadenas, pero estas búsquedas se orientan a la forma de la palabra, nunca al lema. El CdE sí está lematizado y cuenta además con etiquetación morfológica y semántica, lo que multiplica en gran medida las opciones de búsqueda (Davies 2009). No obstante, la información morfológica resulta bastante incompleta en las secciones

del corpus que no corresponden al español contemporáneo. Además, la anotación no siempre resulta lo suficientemente detallada como para poder abordar con garantías el estudio de determinados fenómenos gramaticales (Nieuwenhuijsen 2009, García y Vázquez 2012).

3.4. Control de factores extratextuales

Una última limitación tiene que ver con la consulta de cuestiones que van más allá del texto. La información extratextual de las obras recopiladas en ambos corpus se suele ceñir únicamente a unos pocos aspectos (marco cronológico, procedencia geográfica y género textual), que son tratados en un nivel muy general (por ejemplo, en el CdE la búsqueda cronológica está restringida a un intervalo mínimo de un siglo; y en el CORDE la procedencia geográfica se limita a dar cuenta del país al que pertenece la obra).

Esto impide al investigador controlar toda una serie de factores que pueden tener influencia en las opciones lingüísticas de quien escribe: la caracterización social del autor (y del destinatario), su procedencia dialectal, la intención comunicativa del texto, etc. En consecuencia, resulta complicado realizar estudios de dialectología histórica o de sociolingüística histórica, sencillamente porque no es posible manejar con precisión aquellas variables que son relevantes en estas disciplinas.

4. POST SCRIPTUM

Post Scriptum. Archivo Digital de Escritura Cotidiana es un proyecto desarrollado en la Universidad de Lisboa y cuyo objetivo principal es la creación de un corpus compuesto por 7000 cartas privadas escritas en español y portugués durante la Edad Moderna (3500 cartas por cada lengua) [2]. El proyecto no solo trata de reunir una amplia colección de cartas privadas, sino que también pretende ofrecer su tratamiento filológico en una edición digital online, acompañada de un estudio lingüístico y cultural y de una anotación lingüística del corpus. Los próximos apartados están dedicados a explicar la metodología que se lleva a cabo en Post Scriptum para el tratamiento

de la parte española del corpus (para una explicación general del proyecto véase Vaamonde et al. 2013).

4.1. Búsqueda

Una primera tarea, que está siendo central durante los dos primeros años del proyecto, es la búsqueda y selección de los textos que van a formar parte del corpus. Se trata de localizar 3500 cartas privadas en español que puedan ser contextualizadas y que fuesen producidas entre 1500 y 1833[3]. Por tanto, la primera cuestión que se debe abordar es dónde encontrar este tipo de documentación tan concreta. La respuesta se halla en fondos judiciales e inquisitoriales de diversos archivos históricos de la geografía española.

Estas misivas –en su mayoría inéditas– sobrevivieron de manera excepcional al ser utilizadas por distintos tribunales civiles y religiosos como prueba instrumental de los delitos que estaban siendo juzgados. Los procesos judiciales suelen incluir interrogatorios sociológicos llevados a cabo por inquisidores y jueces varios, lo que permite obtener frecuentemente datos biográficos sobre el autor y/o sobre el destinatario de las cartas procesadas.

En el momento de redactar estas líneas ya se han localizado 2793 cartas, por lo que se espera cerrar la tarea de búsqueda en un plazo relativamente corto. A modo de ejemplo, la figura siguiente muestra el comienzo de una carta manuscrita del siglo XVIII:

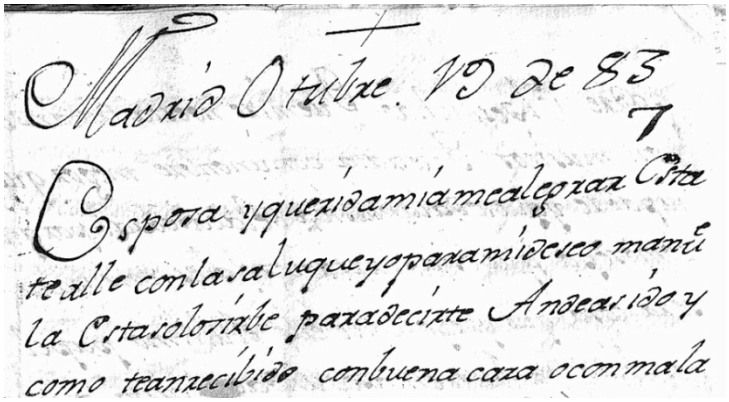

Figura 1. Fragmento de una carta de 1783

4.2. Transcripción

Una vez encontrado el manuscrito, el siguiente paso es transcribirlo, es decir, convertir el texto a un formato que sea legible por un ordenador. Para la codificación de los datos se siguen las directrices de la convención internacional TEI (*Text Encoding Initiative*), que constituye un estándar de referencia en Humanidades Digitales para la edición de textos. Por otro lado, la Definición del Tipo de Documento (DTD) que se utiliza en Post Scriptum es un instrumento proporcionado por el proyecto DALF (*Digital Archive of Letters in Flandes*), dedicado específicamente a la edición epistolográfica[4].

Para la transcripción del manuscrito se ha adoptado una actitud bastante conservadora, lo que da lugar a una edición semipaleográfica del texto original. Tan sólo se ha normalizado la segmentación de palabras y el uso de las grafías "i", "j", "u" y "v". Los cambios de línea, la ortografía, las abreviaturas, las correcciones del autor o los accidentes del soporte, entre otros aspectos, se han respetado en la edición digital. La idea de fondo es la de ofrecer una edición electrónica del texto manuscrito sin perder rigor filológico.

La figura siguiente muestra el fragmento de la carta anterior, transcrito ahora mediante etiquetas TEI-XML:

```
        <opener>
        <salute><deco decoRef="fig1"/></salute>
        <address><addrLine><placeName>Madrid</placeName></addrLine></address>
        <date>Otubre 19 de 83<lb/></date>
        <salute>Esposa y querida mia</salute>
        </opener>
            <p>me alegrar esta<lb/> te alle con la salu que yo para mi deseo
<name>manu<add hand="MS7" place="supralinear">e</add><lb n="false"/>la</name>
esta solo sirbe para decirte Ande as ido y<lb/> como te an recibido con buena
cara o con mala<lb/>
```

Figura 2. Transcripción mediante etiquetas TEI-XML

4.3. Normalización

La variedad ortográfica que presentan los originales se conserva escrupulosamente en la transcripción paleográfica. Sin embargo, esta diversidad se vuelve contraproducente a efectos de tratamiento automático, por lo que se hace necesario un trabajo de normalización del texto como paso previo a su anotación lingüística. En la edición normalizada se presenta la grafía estándar actual de cada palabra, se corrigen aspectos de acentuación y puntuación y se expanden todas las abreviaturas del manuscrito original. Se han conservado los arcaísmos y regionalismos léxicos, marcados con la etiqueta "sic" para facilitar su recuperación. Este proceso se lleva a cabo mediante la herramienta eDictor (Faria, Kepler y Sousa 2010).

La versión normalizada del fragmento anterior quedaría de la siguiente manera (los términos marcados con "sic" se representan aquí en negrita):

Madrid, octubre 19 de 83
Esposa y querida mía.

Me alegrar ésta te halle con la salud que yo para mí deseo.
Manuela, ésta sólo sirve para decirte ***ande*** *has ido y cómo te han recibido, con buena cara o con mala.*

4.4. Anotación

El último paso en el tratamiento del texto es su enriquecimiento lingüístico mediante un anotador automático. Para el caso del español, se está haciendo uso del programa FreeLing (Padró y Stalinovsky 2012) en su versión 3.1, una herramienta de código abierto diseñada para realizar diferentes tareas de análisis lingüístico.

El módulo morfológico de FreeLing asigna a cada palabra su lema correspondiente y una etiqueta morfológica. El etiquetario utilizado se basa en la propuesta del grupo EAGLES para la anotación morfosintáctica de lexicones y corpus para todas las lenguas europeas[5]. Este sistema tiene la ventaja de marcar no sólo la categoría léxica de la palabra (verbo, nombre, ...) sino también las diferentes

categorías flexivas (persona, número, tiempo, modo, ...). Se ofrece a continuación la etiquetación morfológica del ejemplo utilizado a lo largo de todo este apartado (palabra_lema/etiqueta):

Manuela_Manuela/NP00000 ésta_éste/PD0FS000 sólo_sólo/RG sirve_servir/VMIP3S0 para_para/SPS00 decir_decir/VMN0000 te_te/PP2CS000 ande_ande/PT000000 has_haber/VAIP2S0 ido_ir/VMP00SM y_y/CC cómo_cómo/PT000000 te_te/PP2CS000 han_haber/VAIP3P0 recibido_recibir/VMP00SM

4.5. Información extratextual

Todos los subapartados vistos hasta ahora estaban centrados en el tratamiento del texto. Adicionalmente, Post Scriptum ofrece al usuario información extratextual de diferente naturaleza. Entre otras cuestiones, se suelen facilitar los aspectos siguientes:

a) datos contextuales de la carta: fecha, lugar de origen y destino, resumen del contenido, contexto situacional.
b) datos físicos del manuscrito: descripción del soporte, medidas, grafismo, estado de conservación.
c) datos biográficos de los participantes: fecha y lugar de nacimiento, ocupación, parentesco, estado civil, religión, formación, descripción física, etc.

De especial importancia son los detalles biográficos de los participantes, que son organizados y almacenados en una base de datos independiente. Por ejemplo, sabemos que el autor del manuscrito de la Figura 1 se llamaba Manuel Soler, que era hijo de Benito Soler, que era vecino de Valladolid, que en el momento de redactar la carta tenía 23 años y se hallaba en Madrid, que era batidor de oro, que estaba casado con Manuela Herrarte (destinataria de la carta) y que fue denunciado por su suegro por delitos de mala conducta y ociosidad. Toda esta información, obtenida a partir del proceso o de la propia carta y debidamente catalogada, puede ser usada a voluntad del usuario, ya sea con un interés histórico y cultural, ya sea para cruzar con los datos lingüísticos del corpus.

5. CONCLUSIONES

Las características de Post Scriptum, la metodología utilizada en el tratamiento de los textos y la singularidad de los datos recopilados permiten aportar soluciones a las limitaciones de los grandes corpus diacrónicos del español.

La naturaleza dialógica de las cartas privadas, propias de contextos informales, producidas generalmente por manos poco instruidas y escritas casi como si fuesen habladas permiten compensar en su justa medida la carencia de fuentes orales. La triple vía de acceso al documento (transcripción paleográfica, edición normalizada e imagen del facsímil) permite dar respuesta a diferentes perspectivas de estudio sobre los textos, complaciendo así los intereses de un mayor número de usuarios. La anotación lingüística del corpus (en principio en el nivel morfológico, pero en el futuro se pretende dar cabida también al etiquetado sintáctico y semántico) multiplica exponencialmente las opciones de búsqueda, al tiempo que proporciona una base empírica para la extracción de datos lingüísticos sobre diacronía del español. Finalmente, la información extratextual, en especial las fichas biográficas de los participantes, permite realizar búsquedas cruzadas y abre la puerta a estudios de dialectología y sociolingüística históricas.

NOTAS

[1] Existen al menos dos corpus históricos para el español que ofrecen varios niveles de acceso al documento: el proyecto Biblia Medieval, dirigido por Andrés Enrique-Arias, y el proyecto CODEA (Corpus de Documentos Españoles anteriores a 1700), dirigido por Pedro Sánchez-Prieto Borja. El primero es un corpus paralelo de cinco millones de palabras que recoge las traducciones de la Biblia al castellano producidas durante la Edad Media. El segundo consta de 1500 textos, debidamente clasificados y editados, que incluyen documentos cancillerescos, municipales, eclesiásticos y particulares. Ambos recursos son de acceso libre en red.

[2] El proyecto Post Scriptum está siendo financiado por el Consejo Europeo de Investigación (7FP/ERC Advanced Grant - GA 295562).

[3] Se ha adoptado el año 1833 como fecha extrema del corpus, al marcar el inicio del desmantelamiento de las instituciones del Antiguo Régimen con la muerte de Fernando VII.

[4] DALF está basado en el sistema de anotación TEI-P4. En Post Scriptum se está trabajando actualmente en el proceso de conversión a la versión TEI-P5.

[5] En <http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html> se puede acceder a la lista completa de etiquetas EAGLES [comprobado el 09/04/2014].

REFERENCIAS BIBLIOGRÁFICAS

Davies M. 2002. *Corpus del Español: 100 million words, 1200s-1900s*. <http://www.corpusdelespanol.org>

Davies M. 2009. "Creating useful historical corpora: a comparison of CORDE, the Corpus del Español and the Corpus do Português". En Enrique-Arias (ed.) 2009: 137-166.

Enrique-Arias A. (ed.) 2009. *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus*. Madrid/Frankfurt am Main: Iberoamericana/Vervuert.

Enrique-Arias A. 2012. "Dos problemas en el uso de corpus diacrónicos del español: perspectiva y comparabilidad". *Scriptum digital*, 1: 85–106.

Faria P., Kepler F. y de Sousa M. 2010. "An Integrated Tool for Annotating Historical Corpora". *Proceedings of the Fourth Linguistic Annotation Workshop*. 217-221.

García Salido M. y Vázquez Rozas V. 2012: "Los corpus diacrónicos como instrumento para el estudio del origen y distribución de la concordancia de objeto en español". *Scriptum Digital* 1: 67–84.

Nieuwenhuijsen D. 2009. "El rastreo del desarrollo de algunos pronombres personales en español: (im)posibilidades de los corpus diacrónicos digitales". En Enrique-Arias (ed.) 2009: 365-384.

Padró Ll. y Stanilovsky E. 2012. "FreeLing 3.0: Towards Wider Multilinguality". *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA*. Estambul, Turquía. Mayo de 2012

Real Academia Española: Banco de datos (CORDE) [en línea]. *Corpus diacrónico del español.* <http://www.rae.es>

Vaamonde G., Costa A., Marquilhas R., Pinto C. y Pratas, F. 2013. "Philological accuracy and IT resources". *ICHL21: International Conference on Historical Linguistics*, University of Oslo, Oslo, 5-9 de agosto de 2013.